CS591L1 Final Project Report

Alex Lambert, Brian Siao Tick Chong, Chanan Suksangium, Tachapan Kongboonma

Introduction

For our project, our team is curious in finding any form of relationship between a restaurant's Yelp rating and its cleanliness as there are times where we encounter a really good restaurant but the restaurant's condition makes us question their cleanliness. This leads to the formation of our initial hypothesis: Are Yelp ratings of Boston restaurants affected by their health code violations, and if so, are the health code violations and Yelp ratings also affected by the geographical location? In terms of data, we initially used the following four datasets to examine different restaurants and businesses:

1. Code Enforcement - Building and Property Violations (Analyze Boston)
2. Food Establishment Inspections (Analyze Boston)
3. Yelp Fusion
4. Boston Restaurant Licenses (Analyze Boston)

Data is retrieved by calling Yelp Fusion API for each restaurant in Boston Restaurant Licenses dataset, and our group performed data transformation, making it possible to combine the restaurants information retrieved from Yelp containing name, rating, location with our calculated average health violation severity score, found by summing the health severity scores of every health code violation committed by a restaurant and dividing by the total violation count. However, there was lack of matching data for building and property violations, so we could only use datasets 2-4 and thus only take health violations into account. Thus, higher scores suggest that the average restaurants health code violations are more severe (for example: serving raw food would net a higher score compared to utensil cleanliness).

Goal

The list below is our goal for this project:

1. Test whether there is actually any correlation between restaurant's health violation severity and Yelp score.
2. Identify whether restaurants of similar vicinity have the similar health violation scores and Yelp ratings or not.
3. Create web-based service that allows users to find restaurants according to the following features:
    a. Yelp rating
    b. Health violation severity
    c. Number of restaurants
    d. Geographical distance

Analysis

To address the first goal, our group went with Spearman's Rank-Order Correlation primarily because restaurant ratings from Yelp range from 1 to 5 with 0.5 steps between each rank, so Spearman's Correlation made the most sense to deal with ranked data. The statistical analysis found that there is in fact a significant negative correlation between Yelp rating and health violation severity with correlation coefficient of -0.67 and p value of less than 0.05. This shows that there is a negative relationship between restaurant Yelp rating and average health violation. This value can be calculated by running the "BostonRestaurantsStatsAnalysis.py" file.

For the optimization problem, we seek the most ideal location for restaurants by exploring the relationship between rating + safety violation severity and geographical location via longitude and latitude. Initially, k-means clustering was performed on average health violation severity and restaurant rating in hoping to find a distinct clustering as we predicted for lower rated restaurants to be grouped together for higher violation severity and vice versa [see Figure 1].

To address our second goal, restaurants and their relevant information are placed in a graph and distance is calculated based on longitude and latitude. For each restaurant, a score is taken using Yelp rating and average health violation severity (scaled to the same magnitude) to be compared to average score of top 6 closest restaurants; this allows us to compare if geographic locations affect restaurants scores.

Result

For our first goal, to further understand the results we obtain, a visualization was created by mapping the restaurants in Boston [see Figure 1.]. The restaurants are color coded to match with their Yelp rating and severity scores [see Figure 2.]. The colors are chosen by k-means clustering, classified by the rating and safety violation severity score. The visualization showed no observable pockets of commonly scored restaurants. This was the first method our group attempted and found that while there is some correlation between Yelp rating and health violations, there is no correlation with geographic area which contradicts with our second goal of attempting to find an optimal location for a restaurant in Boston.

Figure 1. A mapping of restaurants in Boston   Figure 2. K-Means algorithm plot

Referring back to our second goal, our group aims to identify whether restaurants of similar vicinity have the similar pattern in health violation scores and/or Yelp ratings. Plotting a restaurant's own score compared to the average score of its 6 closest restaurants revealed that geographical location in Boston does not affect a restaurant's score given the insignificant correlation coefficient below 0.2; the three graphs below show the plottings of the average score of 6 closest restaurants vs. each restaurant score. There is no notable correlation between location and all 3 metrics used: health violation score, Yelp rating, or the two combined. This finding limits our next steps because optimization problem of finding the best locations for restaurants in Boston cannot be addressed given the fact that geographical location does not affect a restaurant's Yelp rating nor health violation score. This result can be shown by running "BostonRestaurants_FullyConnectedMap.py" as this file creates distance graph, followed by the file "BostonRestaurantsScoreComparisonAll3.py" to calculate and compare scores.

Our web-based service is the result of our third goal. This is where our group utilized our findings and tried to make it useful for a consumer. This web-based service is built with Flask and LeafletJS on the frontend for visualization. The web-service allows a user to find restaurants based on his/her preference on Yelp rating, health code violation magnitude, and the number of restaurants they want returned [see Figure 3.]. The application utilizes the distance graph and algorithm from analysis step #2 to return and display the best suited restaurants on the map that are reasonably close and meet the requirements set by user.

Figure 3. Screenshot from project 3 web-based service

Conclusion

From a business owner’s perspective, a restaurant will succeed anywhere in Boston since there is no particular area where all the highly rated or cleanest restaurants situate. Focus on cleanliness and safety rather than location for success of your restaurant.

From a consumer’s perspective, there is a negative correlation between health violation score and Yelp rating, suggesting that well-rated restaurants in Boston tend to have good (low) violation scores. Avoid low-rated restaurants not only for quality of food, but also for personal health and safety.

Future Work

Given more time, our team aims to improve the web-based service. One possible option is to include more information when displaying results such as the cuisine of the restaurants, which may potentially suggest what cuisine is known around the particular area of the city and help the user narrow down on restaurant search. Additionally, cuisine could be used as a feature to determine the similarity of restaurants, and so we could perform this analysis for specific types of restaurants rather than all at once. We would also like to extend the findings to other cities, in the US and around the world, to see if Boston is unique, or if the result is consistent with other places as well. Initially analyzing New York City was in the plan, but there was seemingly a lack of matching health violation data on restaurants and the rate limit on Yelp Fusion API made it difficult to scrape the myriad restaurants in New York City.